GENERALIZED SMOOTH FINITE MIXTURES

MATTIAS VILLANI, ROBERT KÖHN, AND DAVID NOTT

ABSTRACT. We propose a general class of models and a unified Bayesian inference methodology for flexibly estimating the density of a response variable conditional on a possibly high-dimensional set of covariates. Our model is a finite mixture of component models with covariate-dependent mixing weights. The component densities can belong to any parametric family, with each model parameter being a deterministic function of covariates through a link function. Our MCMC methodology allows for Bayesian variable selection among the covariates in the mixture components and in the mixing weights. The model’s parameterization and variable selection prior are chosen to prevent overfitting. We use simulated and real data sets to illustrate the methodology.

KEYWORDS: Bayesian inference, Conditional distribution, GLM, Markov Chain Monte Carlo, Mixture of Experts, Variable selection.

1. INTRODUCTION

We propose a general methodology for flexibly estimating the conditional predictive density \( p(y|x) \) for all \( x \), where \( y \) is univariate discrete or continuous, and \( x \) is a possibly high-dimensional set of covariates. Making inferences on the whole density \( p(y|x) \) is crucial in many applications. One obvious example is the Value-at-Risk measure in finance, which is defined as a tail quantile in a return distribution. Another example is the distribution of the aggregate firm default rate in the economy as a function of indicators of macroeconomic activity.

Our article contributes to the field of mixture-of-experts models initiated by Jacobs et al. (1991) and Jordan and Jacobs (1994). The mixture-of-experts models, also named smooth mixture models, are extensions of mixture regression models, such as the clusterwise regression model in DeSarbo and Cron (1988), that allow for covariates in the mixture weights. More specifically, our model is a generalization of the Smooth Adaptive Gaussian Mixture (SAGM) in Villani et al. (2009), which in turn extends the smoothly mixing regression models in Wood et al. (2002) and Geweke and Keane (2007). SAGM is a mixture model with covariate dependent weights modeled as a multinomial logit. The mixture components in the SAGM model are Gaussian with the mean and the log variance being functions of covariates, so SAGM is a model for a continuous response variable. Li et al. (2010) extend the SAGM model by having skewed student-\( t \) components in the mixture. The scope of our model and inference methodology is much larger as the components can belong to essentially any distributional family, so the model applies to both continuous and discrete \( y \), and to densities outside the exponential family. The model for binary data in Wood et al. (2008) is a special case of
the model proposed here. The parameters in the component densities are linked to different sets of covariates via arbitrary link functions. As an example, a smooth mixture of beta-binomial components with both the mean and over-dispersion parameter linked to covariates is a very flexible model for binomial data. Another example is given in Section 5.1 where we model proportions data using a smooth mixture of Beta densities with the mean and the over-dispersion parameter in each component both being functions of covariates. Finally, Section 5.2 models over-dispersed count data using smooth mixtures of Poisson, negative binomial and generalized Poisson components.

A key component of our approach is an extension of the highly efficient MCMC method in Villani et al. (2009). This method generates joint proposals of the model coefficients and the variable selection indicators from a tailored proposal density obtained by taking a few Newton-Raphson steps toward the full conditional mode in each iteration of the algorithm. The model setup with low-dimensional model parameters linked to possibly high-dimensional sets of covariates, and the availability of gradient and Hessians in analytically tractable form, makes this a fast, reliable and very efficient MCMC scheme. Previous literature has developed several clever MCMC schemes for smooth mixtures with specific component densities (Peng et al. (1996); Wood et al. (2002); Wood et al. (2008); Geweke and Keane (2007)). An important advantage of our inferential procedure is that it treats all models in a uniform way, regardless of the distributional form of the component densities. It is therefore possible to write general computer code where completely new component models can be implemented very rapidly. The user only needs to code up the likelihood, gradient and (optionally) the Hessian, but only with respect to the low-dimensional (typically scalar) parameters of the component densities, which in most cases is a trivial exercise.

We also propose a class of priors that are specified using the same principles for any model in the model class. Our completely unified triple of models, priors and algorithm makes it a very convenient framework for a large spectrum of regression problems.

Villani et al. (2009) showed how a smooth mixture of homoscedastic Gaussian components is able to generate some heteroscedasticity, but seldom enough to fit heteroscedastic data well, at least in situations with more than a couple of covariates. We show here that this result seems to generalize to other data types. Using mixtures of over-dispersed components improves the out-of-sample performance of the predictive density compared to smooth mixtures of simpler, equi-dispersed, components. A simulation study shows that when analyzing over-dispersed count data, it is better to allow for over-dispersion in the mixture components than to rely on a smooth mixture of Poissons to generate the over-dispersion, even if the estimated over-dispersed model is mis-specified. An application to German health reform data points clearly in the same direction. As a side result, we find that the less well-known Generalized Poisson regression model in Consul and Jain (1973) and Famoye and Singh (2006) is a strong competitor to the traditionally used Negative Binomial regression model for over-dispersed count data, both in terms of predictive performance, numerical efficiency, and computing time.

Our model class is obviously very flexible and typically over-parametrized. We handle overfitting using three strategies. First, we use a Bayesian approach with easily specified priors on low-dimensional aspects of the model. Second, we use Bayesian variable selection among the
covariates in all the component parameters and in the mixing function. This automatically reduces the number of effective parameters, especially in multi-component models, and also gives insights about the importance of covariates in the different parts of the model. Finally, we parametrize the components as deviations from a reference component. This means that the variable selection prior is not setting coefficients to zero, but rather pushing components toward each other, which can lead to very efficient parsimony in some situations. All in all, our model seems to be very flexible when needed, but is able to simplify when data suggests a more parsimonious data generating process.

There have been several recent articles on modeling conditional densities using Dirichlet process mixture (DPM) models and variants of this model, particularly for the case of a continuous response variable, see e.g. Muller et al. (1996), Dunson et al. (2007) and Chung and Dunson (2009), and a few for discrete responses, see for example Shahbaba and Neal (2009) for multinomial classification. The DPM model is in theory a mixture with an infinite number of components. The main advantages of DPM models are that the number of components is estimated directly from the data, and that theoretical results like posterior consistency are more accessible. The drawback is that the MCMC is both computationally more costly, especially for large data sets, and convergence to the target posterior is slower than for a finite mixture. The reason for the slower convergence is that there is no data available to tailor the proposal for new mixture components, and such components are repeatedly introduced throughout the MCMC iterations. This problem is clearly worse when the mixture components carry many parameters. We also note that DPM models implicitly uses the marginal likelihood for inferences on the number of mixture components which is typically highly sensitive to the models’ priors and to model mis-specification; we use the cross-validated log predictive score which we believe is a more sensible model comparison criteria, see Section 4.3. See also Green and Richardson (2001) for a comparison of finite mixture models and DPM models in the case without covariates.

2. A CLASS OF GENERAL REGRESSION MODELS

This section describes a very general class of regression models that fits naturally into the inference methodology in Section 4. The models in this class will form the components in the smooth mixture described in Section 3.

2.1. The component models. Our component models are general regression densities \( p(y|x) \) that depend on a set of model parameters \( \phi_1, \ldots, \phi_J \). Each parameter is connected to its own set of linear predictors through link functions \( h_j(\phi_j) = \alpha_j + x_j' \beta_j \), where \( \alpha_j \) is the intercept, \( x_j \) is the vector with covariates for the \( j \)th model parameter and \( \beta_j \) is the corresponding regression coefficients. To make a clear distinction between the two different layers of parameters we will consistently use the terminology that the (scalar-valued) model parameters \( \phi_j \) are linked to covariates via the regression coefficients \( \alpha_j \) and \( \beta_j \). The models extend the class of generalized linear models (GLM) in two directions. First, \( p(y|x) \) is not restricted to the exponential family, but can be any well-behaved density satisfying the usual regularity conditions. Second, it allows for covariates in other parameters than the mean, e.g. the variance, skewness or kurtosis. This extends the model in Smyth (1989) who allows for covariates in the mean and the
over-dispersion. One example is the heteroscedastic Gaussian regression model with the two parameters being the mean and the variance, with the identity and the log-link, respectively. Another two-parameter example is the Beta-Binomial model, which is a generalization of the usual Binomial regression model, with response distribution
\[ y|\lambda, \varphi \sim \text{BetaBin}\left[y|\varphi \mu, \varphi \left(1 - \mu \right), n\right], \quad y \in \{0, 1, ..., n\}, \]
where
\[ E(y) = n\mu \quad \text{and} \quad V(y) = n\mu \left(1 - \mu \right) \left(\varphi + 1\right), \]
which shows that \( \varphi \) is the over-dispersion parameter. The two model parameters may be linked to covariates by, for example, \( \ln[\mu/(1 - \mu)] = \alpha_\mu + x'\beta_\mu \) and \( \ln \varphi = \alpha_\varphi + x'\beta_\varphi \). Finally, an example of a three parameter model is the zero-inated negative binomial regression where with probability \( \pi \) we observe \( y = 0 \), and with probability \( 1 - \pi \), \( y \) follows a negative binomial distribution (see Appendix B for details)
\[ y|\mu, \varphi \sim \text{NegBin}\left[y|\varphi \mu + \varphi, \varphi\right], \]
where \( E(y|\mu, \varphi) = \mu \) and \( \text{Var}(y|\mu, \varphi) = \mu \left(1 + \mu \varphi^{-1}\right) \), again showing that \( \varphi \) is an over-dispersion parameter. The three model parameters \( \mu, \varphi \) and \( \pi \), are all linked to covariates through different link functions.

The covariates in the different parameters may be subsets or deterministic functions of the original covariates in \( x \), e.g. spline basis expansions, similar to the generalized additive models in Hastie and Tibshirani (1990) or general surface (interaction) models. Splines will be especially attractive when the parameter is common to all mixture components, see Section 3. Section 4.4 in Villani et al. (2009) has an example with variable selection among the knots in thin plate spline surfaces in the variance and mixing function.

2.2. The priors. Consider a model with two parameters \( \mu \) and \( \varphi \), each connected to its own set of covariates \( v \) and \( w \) via the link functions \( g_\mu(\mu_i) = \alpha_\mu + v_i'\beta_\mu \) and \( g_\varphi(\varphi_i) = \alpha_\varphi + w_i'\beta_\varphi \). The extension to more than two parameters is immediate. We start out by assuming independence between \( \mu \) and \( \varphi \). The prior can be extended to allow for dependence between \( \mu \) and \( \varphi \) by choosing an order of the parameters and eliciting one of the parameters conditionally on the other. Our prior for the regression coefficients is an extension of the prior in Ntzoufras et al. (2003) to the case with more than one parameter, and a density possibly outside the exponential family. Our strategy is to elicit a prior on each parameter (e.g. \( \mu \)) at the mean of the covariates and then derive the implied prior on the intercept (e.g. \( \alpha_\mu \)). The regression coefficients in \( \beta \) are then assigned a Gaussian prior with zero mean and a covariance matrix proportional to the inverse Fisher information evaluated at the prior mode of the model parameters. This prior allows the user to specify prior beliefs on the scalar-valued parameters while the regression coefficients are assigned shrinkage priors based on the inverse Fisher information.

Prior on the intercepts. To elicit the prior on the intercepts, we standardize the covariates by subtracting the mean and dividing by the standard deviation. The intercepts \( \alpha_\mu \) and \( \alpha_\varphi \)
are then simply the model parameters (transformed by the link) at the mean of the original covariates, and we may assume independence between the intercept and the regression coefficients. We assume the intercepts to be independent a priori with distributions $\alpha_{\mu} \sim N(m_{\mu}, s^2_{\mu})$ and $\alpha_{\varphi} \sim N(m_{\varphi}, s^2_{\varphi})$. To specify $m$ and $s^2$ for a given parameter we elicit a suitable prior directly on the parameter and work out the implied mean and variance on the intercept. As an example, with a log link we can elicit a log-normal prior on the parameter with mean $m^*$ and standard deviation $s^*$. The implied prior on the intercept is then $N(m, s^2)$, where
\[
s^2 = \ln \left( \frac{s^*}{m^*} \right)^2 + 1
\]
\[
m = \ln m^* - s^2/2.
\]
For some links it is not possible to solve for $m$ and $s^2$ analytically, but it is straightforward to obtain $m$ and $s^2$ by simulation.

**Prior on the regression coefficients.** Turning now to the prior for the regression coefficients, we will assume priors of the form $\beta_{\mu} \sim N(0, c_{\mu} \Sigma_{\mu})$ and $\beta_{\varphi} \sim N(0, c_{\varphi} \Sigma_{\varphi})$. We now discuss the form of $\Sigma_{\varphi}$; $\Sigma_{\mu}$ is defined analogously. $\Sigma_{\varphi}$ is taken to be the inverse conditional Fisher information
\[
\Sigma_{\varphi} = (W'\hat{D}_{\varphi}W)^{-1},
\]
where $W = (w_1, \ldots, w_n)'$ is the matrix of covariates for $\varphi$ excluding the unity for the intercept, and $\hat{D}_{\varphi}$ is the conditional Fisher information for $\varphi$ evaluated at the prior modes of $\beta_{\mu}$ and $\beta_{\varphi}$ (we use a circumflex to denote a quantity evaluated at the prior mode). Since the prior modes of all parameters except the intercept are zero, $\hat{D}_{\varphi}$ depends only on the two constants $m_{\mu}$ and $m_{\varphi}$. The conditional Fisher information for $\varphi = (\varphi_1, \ldots, \varphi_n)'$ is the $n \times n$ diagonal matrix with elements
\[
-E_{y_i|\mu_i,\varphi_i} \left[ \frac{\partial^2 \ln p(y_i|\mu_i, \varphi_i)}{\partial \varphi_i^2} \right] g_\varphi(\varphi_i)^{-2}.
\]
Setting $c_{\varphi} = n$ gives a unit information prior, i.e. a prior that carries the information equivalent to a single data observation from the model. For some models the conditional Fisher information is not available in closed form, but can then be obtained by straightforward simulation. Note that we only need to perform such a simulation once, before the MCMC, since the prior covariance matrices do not depend on unknown parameters.

It is instructive to examine the details of the proposed prior when the model is a heteroscedastic Gaussian regression with identity link for the mean $\mu$, and log-link for the variance $\varphi$. In this case we have $\hat{D}_{\varphi} = (1/2)I_n$, so
\[
\Sigma_{\varphi} = 2(W'W)^{-1}.
\]
Also, $\Sigma^{-1}_{\mu} = V'Diag(\hat{\varphi}_1^{-1}, \ldots, \hat{\varphi}_n^{-1})V$, which is a sensible choice. In the special case of a homoscedastic regression we obtain the usual shrinkage prior with $\Sigma_{\mu} = \hat{\varphi}(V'V)^{-1}$. The overall prior for $(\alpha_{\mu}, \beta_{\mu})'$ is denoted by $N(\hat{m}_{\mu}, \hat{\Sigma}_{\mu})$, where $\hat{m}_{\mu} = (m_{\mu0}, 0')'$, $0$ is the zero vector, and
\[
\hat{\Sigma}_{\mu} = \begin{pmatrix} s^2_{\mu0} & 0' \\ 0 & \Sigma_{\mu} \end{pmatrix}.
\]
In summary, the user is required to perform the modest task of specifying the prior mean $m^*$ and standard deviation $s^*$ for each scalar-valued parameter (at the mean of the original covariates). The remaining parameters are assigned information priors that automatically take into account the scale of the covariates and the link functions.

**Variable selection.** We will allow for variable selection among the covariates in all parameters. Let $\mathcal{I} \in \{0, 1\}^p$ denote the $p$-dimensional vector with zero-one indicators for the covariates in a given parameter, and $\mathcal{I}^c$ the complement of $\mathcal{I}$. If $\mathcal{I}_i = 0$, then $\beta_i = 0$ and the $i$th covariate in the parameter drops out. The intercept is always included in the model. Let $\beta_\mathcal{I}$ denote the vector with the non-zero coefficients, and note that $\beta_\mathcal{I}$ may be null. Our prior for $\beta_\mathcal{I}$ is the $N(m, \Sigma)$ prior for the full $\beta$ from Section 2.2 conditional on $\beta_\mathcal{I}^c = 0$. This gives us

$$\beta_\mathcal{I} | \mathcal{I} \sim N \left[ 0, c \left( \Sigma_{\mathcal{I},\mathcal{I}} - \Sigma_{\mathcal{I},\mathcal{I}^c} \Sigma_{\mathcal{I}^c,\mathcal{I}}^{-1} \Sigma_{\mathcal{I}^c,\mathcal{I}^c} \right) \right],$$

with obvious notation, and $\beta_{\mathcal{I}^c} | \mathcal{I}$ is identically zero.

We will assume the $\mathcal{I}_i$ to be a priori independent within and between parameters and that $\mathcal{I}_i \sim Bern(\pi)$, where $0 \leq \pi \leq 1$ is the prior inclusion probability, and $\pi$ can be different across the parameters. It is also possible to estimate $\pi$ as in Kohn et al. (2001), or to use a prior on the model size and a uniform distribution on $\mathcal{I}$ for a given model size in Denison et al. (2002). In mixture models with splines, the prior inclusion probability $\pi$ of a knot can be modeled to take into account the mixture weights, see Villani et al. (2009).

Variable selection is important for preventing over-fitting, especially in models with many covariates and for data sets with not so many observations, see for example Figure 1 in Villani et al. (2009) for a nice demonstration. Variable selection is also helpful for MCMC efficiency since it keeps the dimensionality of the proposed coefficient vector to a minimum in every MCMC iteration. This is especially important here as we tailor the proposal to the full conditional posterior.

### 3. Generalized Smooth Mixtures

Our overall model is a finite mixture of component densities $p_k(y|x)$, $k = 1, ..., K$:

$$p(y|x) = \sum_{k=1}^{K} \omega_k(z)p_k(y|x),$$  

with covariate-dependent mixing probabilities $\omega_k(z)$, where $z$ may be subset of the covariates in $x$ or other covariates. Models of the form (3.1) are referred to in the literature as smooth mixtures, or mixture-of-experts models. Following the majority of the literature on mixture-of-experts models, we will assume that the mixing function is a multinomial logit

$$\omega_k(z) = \frac{\exp(z'\gamma_k)}{\sum_{r=1}^{K}\exp(z'\gamma_r)},$$

with $\gamma_1 = 0$ for identification. Following Geweke and Keane (2007) and Villani et al. (2009), we use the rather uninformative $N(0, 10^2 \cdot I)$ prior for $\gamma = (\gamma_2', \ldots, \gamma_K')'$. The component densities $p_k(y|x)$ are models from the general class described in Section 2.
It is interesting to ask if covariates are really needed in both the mixture weights and the component parameters, and much of the early machine learning literature uses very simple experts (see e.g. Jacobs et al., 1991 and Jordan and Jacobs, 1994). Theoretical results in Jiang and Tanner (1999) show that smooth mixtures with simple mixture components can model a fairly large class of densities if the number of mixture components is large enough, but also that the convergence rate toward the true density is quickly deteriorating with the number of covariates. Norets (2010) provides more general conditions under which a smooth mixture of Gaussians can get arbitrarily close to the target density in the Kullback-Leibler distance. An important deficiency of these articles is that they assume no estimation uncertainty, which therefore makes it costless to increase the number of components in the mixture. Simulation studies and real data applications in Villani et al. (2009) show convincingly that the out-of-sample performance can be substantially improved by the use of heteroscedastic Gaussian mixture components, and Section 5 provides additional evidence for cases with a discrete response variable. Put differently, model flexibility is often better handled by increasing the complexity of the components that by generating the flexibility solely from the mixture. Norets (2010) also shows that allowing the components to be heteroscedastic, or heavy-tailed, can weaken the restrictions on the class of the approximable densities. There may of course be cases where different configurations of the smooth mixture model give a similar model fit and predictive distribution, and it may be hard to discriminate between these similar models for a given data set.

We apply the usual reformulation of mixture models to simplify the MCMC inference. Let \( s_1, \ldots, s_n \) be unobserved indicator variables for the observations in the sample such that \( s_i = k \) means that the \( i \)th observation belongs to the \( k \)th component, \( p_k(y|x) \). The model in (3.1) and (3.2) can then be written

\[
\Pr(s_i = k|z_i, \gamma) = \omega_k(z_i) \\
y_i|s_i = k, x_i \sim p_k(y_i|x_i).
\]

Conditional on \( s = (s_1, \ldots, s_n)' \), the mixture model decomposes into \( K \) separate component models \( p_1(y|x), \ldots, p_K(y|x) \), with each data observation being allocated to one and only one component. This property will be exploited in the MCMC in the usual fashion, see e.g. Fruhwirth-Schnatter (2006).

We consider three different parameterizations of the mixture components. Let \( \beta^{(k)} \) denote the regression coefficients of the \( k \)th component for a given parameter. In the separate parameterization the \( \beta^{(k)} \) are independent parameters. In the reference parameterization \( \beta^{(1)} = \beta \) is the reference component, with the remaining components modeled as deviations from this reference component: \( \beta^{(k)} = \beta + \beta^{(k)} \), for \( k = 2, \ldots, K \). Note that the intercepts are not included in the \( \beta \) vectors and the intercepts are always separate coefficients. This is a useful parameterization since variable selection with zero restrictions on \( \beta^{(k)} \) can now be used to simplify the components toward the reference component. In the extreme case \( \beta^{(2)} = \ldots = \beta^{(K)} = 0 \) we are imposing the restriction that the parameter is proportional in the components, e.g. the proportional variance functions in Villani et al. (2009); we call this the proportional parameterization. In addition, variable selection on the reference component can remove a covariate
completely from a given parameter. Note also that we can use different parameterizations on the different parameters in a given model. We will use the same prior from Section 2.2 on the $\beta$ vectors in all the components. It is clearly possible to apply extra prior shrinkage or more restrictive variable selection on $\beta^{(2)}, \ldots, \beta^{(K)}$ compared to $\beta$ in the reference component parameterization, or different shrinkage to different components (e.g. by setting the shrinkage factor in the $k$th mixture component to $c_k = c/k$).

We use the following notation. Let $y = (y_1, \ldots, y_n)'$ be the $n$-vector of responses, and $X = (x_1, \ldots, x_n)'$ the $n \times p_x$ dimensional covariate matrix. Let $X_j = (x_{1j}, \ldots, x_{nj})'$ be the $n \times p_j$ covariate matrix for the $j$th parameter and let $Z$ be the $n \times p_z$ covariate matrix in the mixing function. Let $s = (s_1, \ldots, s_n)'$ denote the $n$-vector of component indicators for the full sample. Furthermore, define the $p_j \times K$ matrix of regression coefficients $\beta_j = (\beta_j^{(1)}, \ldots, \beta_j^{(K)})$ for the $j$th parameter.

4. Inference

4.1. The general MCMC scheme. We use MCMC methods to sample from the joint posterior distribution of the model parameters. The algorithm presented here simultaneously draws the model coefficients and does variable selection. Villani et al. (2009) experimented with several different algorithms in a related setting and the algorithm proposed here is similar to their preferred algorithm.

Our algorithm is a Metropolis-within-Gibbs sampler that draws parameters using the following three main blocks:

1. $s = (s_1, \ldots, s_n)$
2. $\gamma$ and $I_Z$
3. $\{(\beta_j^{(k)}, I_j^{(k)}), \ldots, (\beta_j^{(k)}, I_j^{(k)})\}_{k=1,\ldots,K}$

The first step of the sampler is straightforward since the elements of $s$ are independent conditional on the model parameters and can therefore be drawn simultaneously. Conditional on $s$, Step 2 is a multinomial logistic regression with variable selection. Villani et al. (2009) gives the details on how the general method described in Section 4.2 can be extended to handle this updating step efficiently. Conditional on $s$, Step 3 decomposes into $K$ independent problems, since the parameters in the components are independent conditional on the component allocation $s$. We will sample each $(\beta_j^{(k)}, I_j^{(k)})$ pair from its full conditional posterior using the general method described in the next subsection.


\begin{equation}
(4.1) \quad p(\beta | y) \propto p(y | \beta)p(\beta) = \prod_{i=1}^{n} p(y_i | \phi_i)p(\beta),
\end{equation}
where $k(\phi_i) = x_i' \beta = \eta_i$, $k : \mathbb{R} \rightarrow \mathbb{R}$ is a smooth link function, and $x_i$ is a covariate vector for the $i$th observation. The extension to more than one parameter is immediate. The class of GLMs are of the form in (4.1), but the methodology applies more widely as it is not restricted to the exponential family. We could even have $k(\phi_i) = X_i \beta$ where $X_i$ is a covariate matrix for the $i$th observation and $k(\cdot)$ is a multidimensional link function, but the vector form is sufficient for our purposes here. Note also that $p(\beta|y)$ may be a conditional posterior density and the algorithm can then be used as a step in a Metropolis-within-Gibbs algorithm. The full conditional posterior for each $\beta_1, ..., \beta_J$ (in a given component) are clearly all of the form (4.1). We can now use Newton’s method to iterate $R$ steps from the current point $\beta_c$ in the MCMC sampling toward the mode of $p(\beta|y)$, to obtain $\hat{\beta}$ and the Hessian at $\hat{\beta}$. Note that $\hat{\beta}$ may not be the mode but is typically close to it already after a few Newton iterations since the previously accepted $\beta$ is used as initial value; setting $R = 1, 2$ or $3$ is therefore usually sufficient. This makes the algorithm very fast. The Hessian can also be replaced with its expected value

$$E_{y|\beta} \left[ \frac{\partial^2 \ln p(\beta|y)}{\partial \beta \partial \beta'} \right]$$

in the Newton iterations. This typically improves numerical stability, with only a slightly worse approximation of $p(\beta|y)$.

The following lemma gives compact matrix expressions for the gradient, Hessian and Fisher information of the target likelihood function.

**Lemma 1.** The gradient of the likelihood in (4.1) is of the form

$$\frac{\partial \ln p(y|\beta)}{\partial \beta} = X' g,$$

where $X = (x_1, ..., x_n)'$, $g = (g_1, ..., g_n)'$,

$$g_i = \frac{\partial \ln p(y_i|\phi_i)}{\partial \phi_i} \left[ k'(\phi_i) \right]^{-1},$$

and the Hessian is

$$\frac{\partial^2 \ln p(y|\beta)}{\partial \beta \partial \beta'} = X' \left[ \text{Diag}(d_{1i}) + \text{Diag}(d_{2i}) \right] X$$

where

$$d_{1i} = \frac{\partial^2 \ln p(y_i|\phi_i)}{\partial \phi_i^2} \left[ k'(\phi_i) \right]^{-2}$$

and

$$d_{2i} = -\frac{\partial \ln p(y_i|\phi_i)}{\partial \phi_i} k''[k'(\phi_i)^{-1}]k'(\phi_i)^{-2}.$$
If the prior for $\beta$ is $N(\mu_\beta, \Sigma_\beta)$ then the prior gradient and Hessian are
\[
\frac{\partial \ln p(\beta)}{\partial \beta} = -\Sigma_\beta^{-1}(\beta - \mu_\beta) \quad \text{and} \quad \frac{\partial^2 \ln p(\beta)}{\partial \beta \partial \beta'} = -\Sigma_\beta^{-1}.
\]

Having obtained good approximations of the posterior mode and covariance matrix from the Newton iterations, the proposal $\beta_p$ is now drawn from the multivariate $t$-distribution with $v > 2$ degrees of freedom:
\[
\beta_p | \beta_c \sim t \left[ \tilde{\beta}, -\left( \frac{\partial^2 \ln p(\beta | y)}{\partial \beta \partial \beta'} \right)_{\beta = \tilde{\beta}}, v \right],
\]

where the second argument of the density is the covariance matrix and $\tilde{\beta}$ is the terminal point of the $R$ Newton steps.

Consider now the case of variable selection. The $p$-dimensional parameter vector $\beta$ is then augmented by a vector of binary covariate selection indicators $I = (i_1, ..., i_p)$, and $\beta$ and $I$ are proposed simultaneously using the decomposition
\[
g(\beta_p, I_p | \beta_c, I_c) = g_1(\beta_p | I_p, \beta_c)g_2(I_p | \beta_c, I_c),
\]

where $g_2$ is the proposal distribution for $I$ and $g_1$ is the proposal density for $\beta$ conditional on $I_p$. The Metropolis-Hasting acceptance probability is
\[
a[\beta_c, I_c \rightarrow (\beta_p, I_p)] = \min \left( 1, \frac{p(y | \beta_p, I_p)p(\beta_p | I_p)p(I_p)p(\beta_c | I_c, \beta_p)g_2(I_c | \beta_p, I_p)}{p(y | \beta_c, I_c)p(\beta_c | I_c)p(I_c)p(\beta_p | I_p, \beta_c)g_2(I_p | \beta_c, I_c)} \right).
\]
The proposal density at the current point $g_1(\beta_c | I_c, \beta_p)$ is a multivariate $t$-density with mode $\tilde{\beta}$ and covariance matrix equal to the negative inverse Hessian evaluated at $\tilde{\beta}$, where $\tilde{\beta}$ is the point obtained by iterating $R$ steps with the Newton algorithm, this time starting from $\beta_p$. A simple way to propose $I_p$ is to randomly select a small subset of $I_c$ and then always propose a change of the selected indicators (Metropolized move), see Nott and Kohn (2005) for other proposals. It is important to note that $\beta_c$ and $\tilde{\beta}$ may now be of different dimensions, so the original Newton iterations no longer apply. We will instead generate $\beta_p$ using the following generalization of Newton’s method. We exploit the idea that when the parameter vector $\beta$ changes dimensions, the dimension of the parameter $\varphi_c = k^{-1}(x'\beta_c)$ and $\varphi_p = k^{-1}(x'\beta_p)$ stay the same, and the two functionals are expected to be quite close. A generalized Newton update is
\[
(4.4) \quad \beta_{r+1} = A_r^{-1}(B_r \beta_r - g_r), \quad (r = 0, ..., R - 1),
\]

where $\beta_0 = \beta_c$, and the dimension of $\beta_{r+1}$ equals the dimension of $\beta_p$, and
\[
g_r = X_{r+1}'d + \frac{\partial \ln p(\beta)}{\partial \beta} \\
A_r = X_{r+1}'(D_1 + D_2)X_{r+1} + \frac{\partial^2 \ln p(\beta)}{\partial \beta \partial \beta'} \\
B_r = X_{r+1}'(D_1 + D_2)X_r + \frac{\partial^2 \ln p(\beta)}{\partial \beta \partial \beta'},
\]
where $X_r$ is the covariate matrix corresponding to the model with $\beta = \beta_r$, and all expressions are evaluated at $\beta = \beta_r$.

After the first Newton iteration the parameter vector no longer changes in dimension, and the generalized Newton algorithm in (4.4) reduces to the original Newton algorithm. The proposal density $g_1(\beta_p|I_p, \beta_c)$ is again taken to be the multivariate $t$-density in exactly the same way as in the case without covariate selection. Once the simultaneous update of the $(\beta, I)$-pair is completed, we make a final update of the non-zero parameters in $\beta$, conditional on the previously accepted $I$, using the fixed dimension Newton algorithm.

It is important to note that it is only necessary to compute the scalar derivatives $\partial \ln p(y_i|\phi_i)/\partial \phi_i$ and $\partial^2 \ln p(y_i|\phi_i)/\partial \phi_i^2$ and the first two derivatives of the link function. This makes it easy to write general computer code that can be adapted to many different distributions and link functions.

In the reference component parameterization (see Section 2.1), the parameters of the reference component appear in all of the $K$ components. The structure of the updating steps are the same, however, the only difference is that the number of elements in the $g$ vector in Lemma 1 is the same as the full sample, not just the observations currently allocated to the reference component. Each element $g_i$ is still computed with respect to the component to which observation $i$ is currently allocated to. The same applies to the Hessian.

### 4.3 Model selection

The number of components is assumed known in the MCMC scheme above. A Bayesian analysis of mixture models with an unknown number of components is possible using e.g., Dirichlet process mixtures (Escobar and West, 1995), reversible jump MCMC (Richardson and Green, 1997) and birth-and-death MCMC (Stephens, 2000). The fundamental quantity determining the posterior distribution of the number of components is the marginal likelihood of the models with different number of components. It is well-known, however, that the marginal likelihood is sensitive to the choice of prior, and this is especially true when the prior is not very informative, see e.g. Kass (1993) for a general discussion and Richardson and Green (1997) in the context of density estimation.

Following Geweke and Keane (2007) and Villani et al. (2009), we therefore compare and select models based on the cross-validated Log Predictive Score (LPS). By sacrificing a subset of the observations to update/train the vague prior we remove much of the dependence on the prior, and obtain a better assessment of the predictive performance that can be expected for future observations. To deal with the arbitrary choice of which observations to use for estimation and model evaluation, we use $B$-fold cross-validation of the LPS:

$$
\frac{1}{B} \sum_{b=1}^{B} \ln p(\tilde{y}_b|\tilde{y}_{-b}, x),
$$

where $\tilde{y}_b$ is an $n_b$-dimensional vector containing the $n_b$ observations in the $b$th test sample and $\tilde{y}_{-b}$ denotes the remaining observations used for estimation. If we assume that the observations are independent conditional on the parameters $\theta$, then

$$
p(\tilde{y}_b|\tilde{y}_{-b}, x) = \int \prod_{i \in T_b} p(y_i|\theta, x_i)p(\theta|\tilde{y}_{-b}) d\theta
$$
where $T_b$ is the index set for the observations in $\tilde{y}_b$, and the LPS is easily computed by averaging $\prod_{i \in T_b} p(y_i|\theta, x_i)$ over the posterior draws from $p(\theta|\tilde{y}_{-b})$. This requires sampling from each of the $B$ posteriors $p(\theta|\tilde{y}_{-b})$ for $b = 1, \ldots, B$, but these MCMC runs can all be run independently from each other and are therefore ideal for straightforward parallel computations on widely available multi-core processors. Our machines have four processing cores, so we set $B = 4$ in our applications. Cross-validation is less appealing in a time series setting since it is typically false that the observations are independent conditional on the model parameters for time series data. A more natural approach is to use the most recent observations in a single evaluation sample (Geweke and Keane, 2007).

5. Applications


Our data set is similar to the one in Cook et al. (2008). The data are taken from the Compustat database and covers 4405 American non-financial firms with positive sales in 1992 and complete data records. We analyze the leverage in terms of total debt as in Rajan and Zingales (1995), rather than focusing on long-term debt as in Cook et al. (2008). The variables market2book and profit contain some substantial outliers. Instead of removing these outliers completely, we Winsorize (i.e. setting all tail values to a specified quantile) market2book value at the 99.5 percentile and profit at the 0.5 and 99.5 percentiles. Figure 5.1 plots leverage (in original form to the left and logit transformed to the right) against each of the four covariates. The relationships between the leverage and the covariates are clearly highly non-linear, and the logit transformation of leverage is unable to linearize its relationship with market2Book and profit. The scatter for profit is especially peculiar, suggesting quite different relations between leverage and profits when profits are positive (strong negative relation) from when profits are negative (constant or slightly positive relation). Strong non-linearities seem to be a quite general feature of balance sheet data, and there has been a few very recent attempts with nonlinear/nonparametric models for these data, see Jacobson et al. (2011) for an additive spline approach and Bastos and Ramalho (2010) for an application with regression trees.

Another common feature of balance sheet data is that the response variable is often a proportion, i.e. restricted to the unit interval. This has until very recently been essentially ignored in the empirical literature, but recent modeling efforts are taking this into account, as shown by the recent survey in Ramalho et al. (2009) on fractional regression models; see also Casarin et al. (2010) for a Bayesian analysis of the Beta autoregressive process. We introduce a smooth mixture of Beta components

$$y|\mu, \varphi \sim \text{Beta} [\mu \varphi, (1 - \mu) \varphi],$$
Figure 5.1. Scatter plots of the leverage data with superimposed linear regression lines. The graphs to the left display the leverage on the original scale, whereas the graphs to the right plot the logit transformed leverage.

where $0 \leq \mu \leq 1$ is the mean and $\varphi > 0$ is the precision parameter. The variance of $y$ in this parameterization is $\mu(1-\mu)/(1+\varphi)$. We use a logit link for $\mu$ and log link for $\varphi$

$$
\ln \frac{\mu}{1-\mu} = \alpha_\mu + x^T \beta_\mu
$$

$$
\ln \varphi = \alpha_\varphi + x^T \beta_\varphi.
$$
The prior on $\alpha_\mu$ is set so that the prior mean of $\mu$ equals the sample mean 0.367, but with an implied prior standard deviation on $\mu$ of 0.2, so the prior is rather uninformative. The prior mean of $\alpha_{\phi}$ is set so that components' variances matches the sample variance, but the prior standard deviation of $\phi$ is 10, so the prior is again uninformative.

Previous literature used only a single Beta component without covariates in $\phi$. To illustrate how the smooth mixture of Beta components captures the substantial nonlinearities in the data, Figure 5.2 shows a scatter plot of leverage against the covariate profit and Highest Predictive Density (HPD) regions for the smooth mixture models with 1 to 3 components. Our smooth mixture model is non-additive and we therefore plot the fit from models using only profit as an explanatory variable to simplify interpretation. We also show the predictive HPD intervals from single-component Beta regression with ten thin plate spline knots in the mean function. The spline model is estimated using Bayesian variable selection of the knots. Figure 5.2 shows that a one-component Beta regression model without spline terms has no chance of fitting the data well. The models with two and three Beta components do a much better job and show an interesting distinct negative relation between leverage and profit when profit is positive. This result, which is completely missed by the one-component model, is also present in models with more than three components, and to some extent also in the spline model.

Turning now to the full model with all four covariates, Table 1 presents the four-fold cross-validated out-of-sample Log Predictive Score (LPS) for the smooth mixture of Beta models with 1-5 components. The first three rows correspond to one of the three different mixture parameterizations. As a reference, the last row of Table 1 gives the LPS for a Beta model with ten additive thin plate spline knots in the mean for each covariate. The logarithm of the precision parameter $\phi$ is linear in the covariates. We use Bayesian variable selection among the knots with a mixture of normals prior, $0.5 \cdot I_0 + 0.5 \cdot \phi(0, 10, I)$, where $I_0$ is the Dirac measure on the point zero, on the spline coefficients. The most striking result from Table 1 is how much inferior a one-component Beta model is to the multi-component models. This result is consistent with the very poor in-sample fit of the one-component model in Figure 5.2, and is something that casts doubt on most of the previous analyses on similar data sets. The use of additive splines gives a substantial improvement in LPS compared to the one-component model, but is still much inferior to the multi-component models. It is straightforward to estimate multi-component models with splines, see Villani et al. (2009) for a useful variable selection prior especially tailored for mixtures of splines, but this is not pursued here. It is clear from Table 1 that at least three, probably four, and maybe even five components are needed. It is interesting to observe that the reference parameterization does very well in a situation where its centering model (the model with a proportional scale parameter) seems to be too restrictive, i.e. the components in the reference parameterization can be separated whenever the data requires it.

5.2 Health reform data. Our next data set is taken from the German Socioeconomic Panel (GSOEP, see e.g. Riphahn et al. (2003)) and consists of a patient survey with responses from 2227 German women regarding the number of visits to a physician during the most recent
Figure 5.2. The first two rows depict the 68% and 95% Highest Predictive Density regions from the smooth mixture of Beta models with 1-3 components and a Beta regression model with thin plate spline terms in the mean. The two graphs in the last row both display a data scatter of Leverage vs Profit to aid in assessing the models’ fit.

Table 1. Firms’ leverage data. Four-fold cross-validated log predictive scores (LPS).
three month period. One of the goals of the data collection was to investigate if the number of visits to a physician had declined as a result of the 1997 health reform. We will use the seven covariates from the preferred model in Hilbe (2007): reform (0 = interviewed before reform, 1 = interviewed after reform), badh (0 = good health, 1 = bad health), age (two dummy variables for the age groups 40-49 years and 50-60 years. The group 20-39 years is the reference), educ (two dummy variables for 10.5-12, and 12- years of education, respectively. The reference group is 7-10.5 years of education), loginc (log of household income in DM).

Hilbe (2007) concludes that the data are clearly over-dispersed. As a result, the Poisson regression model does not fit the data well, whereas the negative binomial regression does a much better job. We will evaluate the performance of smooth mixtures of Poisson components, smooth mixtures of negative binomial components and also smooth mixtures of the generalized Poisson model in Consul and Jain (1973).

The negative binomial regression model (NB) is of the form

\[ y \mid \mu, \varphi \sim \text{NegBin} \left[ y \mid \frac{\varphi}{\mu + \varphi} \right], \]

with link functions

\[
\ln \mu = x'\alpha \\
\ln \varphi = x'\delta,
\]

where \( \mu \) is the mean and \( \varphi \) is the over-dispersion parameter. The variance is \( \mu(1 + \varphi^{-1}\mu) \). The Poisson model is the limiting special case when \( \varphi \to \infty \).

We will also consider another over-dispersed generalization of the Poisson model: the generalized Poisson (GP) introduced in Consul and Jain (1973). We use the parameterization in Famoye and Singh (2006) with density

\[
p(y \mid \mu, \varphi) = \frac{1}{y!} \left( \frac{\mu}{1 + \varphi\mu} \right)^y (1 + \varphi\mu)^{y-1} \exp \left[ -\mu \left( \frac{1 + \varphi y}{1 + \varphi \mu} \right) \right],
\]

where \( \mu \) is the mean and \( \varphi \) is the dispersion parameter. The variance of a GP(\( \mu, \varphi \)) variable is \( \mu(1 + \varphi \mu)^2 \). The GP is under-dispersed when \( \varphi < 0 \), but we will restrict \( \varphi \) to be non-negative since allowing for under-dispersion introduces a restriction on \( \varphi \) which is difficult to enforce when we allow for covariates in \( \varphi \), see also Czado et al. (2007). In any case, under-dispersion is rarely encountered in applications, and certainly not in our application here. Note that the usual Poisson is obtained when \( \varphi = 0 \). Log links are used for both \( \mu \) and \( \varphi \). We are not aware of any previous Bayesian implementations of the Generalized Poisson regression model.

We let all seven covariates and a constant enter in all parameters and also in the mixing function. The prior mean on the intercept in \( \mu \) is set so that the model’s expectation equals the sample mean, but with a prior standard deviation of 10 to minimize the effect of data-snooping. Similarly, the prior mean of the intercepts in \( \varphi \) in the NB and GP models are set such that the implied variances of the components equal the sample variance. The prior standard deviation of the intercept in \( \varphi \) is set to 10 in the NB model and to unity in the GP model. This gives fairly non-informative priors in both models. Note that the scale of the
over-dispersion parameter is different in the two models. This can be seen by comparing the model variances: \( \mu(1 + \varphi^{-1}\mu) \) in the NB model versus \( \mu(1 + \varphi\mu)^2 \) in the GP model.

A common way of assessing the fit of a count data model is by comparing the model’s implied marginal distribution of the response counts to the observed counts in the data, see e.g. Hilbe (2007). The left panel of Figure 5.3 displays the fit of the smooth mixture of Poissons model for 1-4 components. It is clear that the models with one or two components generate distributions that do not resemble the data, and it seems to take at least four components to capture the marginal distribution of the response. The right panel of Figure 5.3 shows that once we allow for over-dispersed components, a single component seems to be sufficient to fit the data. Adding more over-dispersed components gives essentially the same marginal distribution of counts (not shown). Note however that the evaluation in Figure 5.3 is in-sample and does not show if the models are generating accurate conditional distributions \( p(y|x) \) out-of-sample. We will now present a more detailed evaluation of out-of-sample performance of the models.

Table 2 displays the four-fold cross-validated log predictive score (LPS) for the three models: Poisson, NB and GP with 1-5 mixture components. The four cross-validation samples are sampled systematically from the original data set. Table 2 reports the LPS for NB and GP when the over-dispersion parameter follows the separate, proportional and reference parameterization, see Section 3. Looking first at the one-component models, it is clear from Table 2 that the out-of-sample performance of the Poisson model is dramatically inferior to the two over-dispersed models, and that GP is almost two LPS units better than NB. The Poisson model improves as more mixture components are added, but even the best smooth mixture of Poissons (the five-component model) is more than 6 LPS units worse than the best over-dispersed mixture (NB with three separate components). The three parameterizations perform similarly.

In each partitioning of the data set, the LPS evaluation in Table 2 uses three quarters of the data (1670 observations) to estimate the models and only one quarter (557 data points) for evaluation. Table 3 presents results for the opposite setup where one quarter in each
Table 2. Cross-validated LPS for the models of the health reform data. Four-fold cross-validation using 75% of the observations for estimation in each partition and the remaining 25% for evaluation.

<table>
<thead>
<tr>
<th>Component models</th>
<th>No. of comp.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Poisson</td>
<td>-1495.16</td>
</tr>
<tr>
<td>Negative binomial - Separate $\varphi$</td>
<td>-1143.46</td>
</tr>
<tr>
<td>Negative binomial - Proportional $\varphi$</td>
<td>-1138.43</td>
</tr>
<tr>
<td>Negative binomial - Reference $\varphi$</td>
<td>-1138.29</td>
</tr>
<tr>
<td>Generalized Poisson - Separate $\varphi$</td>
<td>-1141.62</td>
</tr>
<tr>
<td>Generalized Poisson - Proportional $\varphi$</td>
<td>-1139.61</td>
</tr>
<tr>
<td>Generalized Poisson - Reference $\varphi$</td>
<td>-1141.42</td>
</tr>
</tbody>
</table>

Table 3. Cross-validated LPS for the models of the health reform data. Four-fold cross-validation using 25% of the observations for estimation in each partition and the remaining 75% for evaluation.

<table>
<thead>
<tr>
<th>Component models</th>
<th>No. of comp.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Poisson</td>
<td>-4491.95</td>
</tr>
<tr>
<td>Negative binomial - Separate $\varphi$</td>
<td>-3431.86</td>
</tr>
<tr>
<td>Negative binomial - Proportional $\varphi$</td>
<td>-3422.13</td>
</tr>
<tr>
<td>Negative binomial - Reference $\varphi$</td>
<td>-3422.46</td>
</tr>
<tr>
<td>Generalized Poisson - Separate $\varphi$</td>
<td>-3427.97</td>
</tr>
<tr>
<td>Generalized Poisson - Proportional $\varphi$</td>
<td>-3421.16</td>
</tr>
<tr>
<td>Generalized Poisson - Reference $\varphi$</td>
<td>-3424.72</td>
</tr>
</tbody>
</table>

Partitioning is used for estimation and three quarters are used for evaluation. This is a useful experiment for determining how the different models and parameterizations perform in situations with small (or noisy) estimation data sets. The mixture of Poisson models is now much poorer in comparison to the over-dispersed mixtures, even if we discount for the fact that discriminatory power between the models is larger since much more data are now reserved for evaluation. The best mixture of Poissons (the four-component model) is now 76 LPS units worse than the best over-dispersed mixture (GP with two proportional components). With a smaller proportion of the data used for estimation, the restriction to proportional over-dispersion parameters outperforms the model with separate components. Interestingly, the reference and proportional parameterizations have very similar performance, suggesting that the reference parameterization is being efficiently pruned toward the model with proportional $\varphi$-functions.

Table 4 reports summaries of the posterior inferences for the one-component models. The posterior mean of the regression coefficients in $\mu$ are very similar across models. The reform variable has a coefficient of the expected sign, but it does not enter the mean significantly (here taken to mean that the posterior inclusion probability is larger than 0.5) in the two over-dispersed models, whereas it has a posterior inclusion probability of unity in the Poisson model. This is a reflection of the well known spurious significance of covariates obtained when equi-dispersed models are fitted to over-dispersed data, see e.g. Hilbe (2007). As expected,
bad health clearly leads to an increased number of visits in all three models; in the GP model, 
badh is also significant in the over-dispersion parameter. Note that the dispersion parameter 
has a very different interpretation in the NB model (large $\varphi$ means small over-dispersion) and 
the GP model (large $\varphi$ means large over-dispersion), so their estimates cannot be compared.

The MH acceptance probabilities are very high, and the mixing of the MH chains are 
excellent: the mean inefficiency factors are 5.04 (Poisson), 2.49 (NB) and 1.76 (GP). Total 
computing time for four-fold cross-validation of the Poisson mixture with all 1–5 components is 
26 minutes using uncompiled Matlab 2010b code on a quad-core Intel i7-960 processor running 
64-bit Linux. The corresponding computing time for the NB and GP models (with separate 
components) are 95 and 59 minutes, respectively. The excellent numerical performance of the 
GP model and the fact that it is much faster to estimate than the negative binomial, makes 
it a very interesting model for Bayesian regression analysis of over-dispersed count data.

5.3. Simulation study: Is a smooth mixture of Poissons a good model for over-
dispersed data? The results in Villani et al. (2009) show that smooth mixtures of ho-
moechedastic Gaussian regressions may not perform well on heteroscedastic Gaussian data, 
especially when there are more than a couple of covariates. The health reform example in the
previous section showed that a smooth mixture of a fairly large number of Poisson components is outperformed by over-dispersed one-component models on over-dispersed data. We conduct a small simulation study to further explore this issue with data from the following data generating process (DGP):

$$y_i | x_i, \mu_i, \varphi_i \sim \text{NegBin}(\mu_i, \varphi_i)$$

$$\ln \mu_i = 2$$

$$\ln \varphi_i = 1 - 0.5x_1 + 0.5x_2 - \ldots - 0.5x_q.$$ We consider four values for $q$: $q = 1, 3, 5$ and $10$, to see how the results depend on the dimensionality of the covariate space. The covariates are iid $N(0, 1)$ variables. We generate 50 data sets, each with 1000 observations from the negative binomial regression above. On each data set we fit a smooth mixture of Poissons with $1 - 5$ components and compare the results to an estimated negative binomial regression. We also include a (one-component) generalized Poisson model in the set of estimated models to see how a mis-specified over-dispersed model performs in comparison to the Poisson mixture. Note that the mean of the DGP does not include any active covariates, but such covariates are allowed for in the estimated models, and we use variable selection. The prior on the intercept of the components are mildly data-based and centered around the sample mean and standard deviation, but is fairly non-informative with a standard deviation of 10.

The performance of the models is measured by the expected Kullback-Leibler (KL) divergence between the predictive density of a given estimated model $q(y|x)$ and the DGP $p(y|x)$

$$\text{KL}(p, q) = \int \left[ \int \ln \left( \frac{p(y|x)}{q(y|x)} \right) p(y|x) dy \right] p(x) dx,$$

where $p(x)$ is marginal distribution of the covariates in the DGP. The integral with respect to $y$ is approximated numerically over a fine grid, and the integral with respect to $x$ is approximated by taking the average over 500 random points from the DGP (the same 500 points are used in all estimated models). To control for the sampling variability we report our results as log KL ratios

$$\kappa_q = \ln \frac{\text{KL}(p, q)}{\text{KL}(p, h)},$$

where $h(y|x)$ is the predictive density of the estimated NB model. A value of $\kappa_q$ larger than zero means that $q(y|x)$ is farther away from the DGP than the estimated NB model.

Figure 5.4 summarizes the simulations results, with each panel displaying box plots of the log KL ratios in (5.2) for a given number of covariates in the DGP. Figure 5.4 shows that the estimated NB model does a lot better that a smooth mixture of a Poisson components, even when the number of components is large. The improvement obtained by adding more Poisson components seems to level off and even with five components it is clearly inferior to the NB, but also to the GP model. Figure 5.4 also shows that the relatively poor performance of the mixture of Poissons model becomes worse when the number of covariates grows. The median log KL ratio for a smooth mixture of five Poisson components is 2.175, 2.355, 2.510 and 2.575, with 1, 3, 5 and 10 covariates, respectively.
Figure 5.4. Fitting a smooth mixture of Poisson components to over-dispersed negative binomial data. Each subgraph displays box plots of the log KL ratio (Eq. 5.2) over the 50 simulated data sets. A log KL ratio larger than zero indicates that the predictive density of the estimated $NB(1)$ is closer to the true predictive density than the model under consideration. Results from the estimated $GP(1)$ are also shown in the graphs.

6. Conclusions

We introduce a flexible smooth finite mixture model for modeling the conditional distribution of a continuous or discrete response variable. The components of the mixture can belong to essentially any distributional family with its parameters linked with smooth link functions to possibly high-dimensional covariates. The models are estimated with a unified MCMC methodology that simultaneously draws the regression coefficients and does variable selection in all parameters of the mixture components and in the mixing weights. Applications on two real data sets with counts and proportions as response variables show that the models are interesting flexible alternatives to traditionally used models in the economics, finance and other social sciences. A small simulation study corroborates the conclusion from one of the applications that smooth mixtures of over-dispersed models tend to outperform smooth mixtures of equi-dispersed components, even when the number of equi-dispersed components is large. Our use of Bayesian variable selection to prune the components toward a reference component is shown to give the right parsimony when data suggest similar components, without sacrificing flexibility when non-similar components are needed.
References


Appendix A. Proof of Lemma 1

Proof. The gradient is
\[
\frac{\partial \ln p(y|\beta)}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial \ln p(y_i|\phi_i)}{\partial \phi_i} \frac{\partial \phi_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial \ln p(y_i|\phi_i)}{\partial \phi_i} [k'(\phi_i)]^{-1} x_i.
\]

The Hessian is of the form
\[
\frac{\partial^2 \ln p(y|\beta)}{\partial \beta \partial \beta'} = \sum_{i=1}^{n} \frac{\partial^2 \ln p(y_i|\beta)}{\partial \beta \partial \beta'}
\]
where
\[
\frac{\partial^2 \ln p(y_i|\beta)}{\partial \beta \partial \beta'} = x_i \left(\frac{\partial^2 \ln p(y_i|\phi_i)}{\partial \phi_i^2} \frac{\partial \phi_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta} + \frac{\partial \ln p(y_i|\phi_i)}{\partial \phi_i} \frac{\partial^2 \phi_i}{\partial \eta_i^2} \frac{\partial \eta_i}{\partial \beta} \right) + x_i \left(\frac{\partial \ln p(y_i|\phi_i)}{\partial \phi_i} \frac{\partial \phi_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta} \right) \frac{\partial \ln p(y_i|\phi_i)}{\partial \phi_i} \frac{\partial^2 \phi_i}{\partial \eta_i^2} \frac{\partial \eta_i}{\partial \beta} \right)
\]
\[
= x_i \left(\frac{\partial^2 \ln p(y_i|\phi_i)}{\partial \phi_i^2} \left(\frac{x_i}{[k'(\phi_i)]^{-2} - \frac{\partial \ln p(y_i|\phi_i)}{\partial \phi_i} k'' [k'(\phi_i)^{-1} k'(\phi_i)^{-2}] x_i'}
\right)
\]

The expected Hessian follows from (4.3) and that the score vector \(\partial \ln p(y_i|\phi_i)/\partial \phi_i\) has zero expectation under the usual regularity conditions.

\[\square\]

Appendix B. Details for the negative binomial regression

The density for a single observation \(y|\mu, \varphi \sim NegBin\left(y|\frac{\varphi}{\mu+\varphi}, \varphi\right)\) is of the form
\[
p(y|\mu, \varphi) = \frac{\Gamma(y+\varphi)}{\Gamma(\varphi) y!} \left(\frac{\mu}{\mu+\varphi}\right)^y \left[\frac{\varphi}{\mu+\varphi}\right]^\varphi.
\]

We then have
\[
\frac{\partial \ln p(y|\mu, \varphi)}{\partial \mu} = \frac{y}{\mu} - \frac{y + \varphi}{\mu + \varphi},
\]
\[
\frac{\partial^2 \ln p(y|\mu, \varphi)}{\partial \mu^2} = -\frac{y}{\mu^2} + \frac{y + \varphi}{(\mu + \varphi)^2}
\]
so that
\[
E_{y|\mu, \varphi} \left(\frac{\partial^2 \ln p(y|\mu, \varphi)}{\partial \mu^2}\right) = -\frac{1}{\mu} + \frac{1}{\mu + \varphi}.
\]
Also,
\[
\frac{\partial \ln p(y|\mu, \varphi)}{\partial \varphi} = 1 + \ln \varphi - \ln(\mu + \varphi) - \kappa_0(\varphi) + \kappa_0(y + \varphi) - \frac{y + \varphi}{\mu + \varphi}
\]
\[
\frac{\partial^2 \ln p(y|\mu, \varphi)}{\partial \varphi^2} = \kappa_1(y + \varphi) - \kappa_1(\varphi) + \frac{\mu^2 + \varphi y}{\varphi(\mu + \varphi)^2},
\]
and
\[
E_{y|\mu, \varphi} \left( \frac{\partial^2 \ln p(y|\mu, \varphi)}{\partial \varphi^2} \right) = E_{y|\mu, \varphi}(\kappa_1(y + \varphi)) - \kappa_1(\varphi) + \frac{\mu}{\varphi(\mu + \varphi)},
\]
where \(\kappa_0(\cdot)\) and \(\kappa_1(\cdot)\) are the digamma and trigamma functions, and \(E_{y|\mu, \varphi}(\kappa_1(y + \varphi))\) is easily obtained by simulation.

The gradient and Hessians may be numerically unstable for large \(\varphi\). The following asymptotic formulas may then be useful
\[
\lim_{\varphi \to \infty} \varphi^2 \frac{\partial \ln p(y|\mu, \varphi)}{\partial \varphi} = -\left[ (y - \mu)^2 - y \right] / 2
\]
\[
\lim_{\varphi \to \infty} \varphi^3 \frac{\partial \ln p(y|\mu, \varphi)}{\partial \varphi} = \left[ (y - \mu)^2 - y \right].
\]
We suggest to use these asymptotic expressions for the gradient and Hessian whenever \(\varphi > 10\mu\).

**Appendix C. Details for the Generalized Poisson regression**

The generalized Poisson PDF is given in (5.1). We then have
\[
\frac{\partial \ln p(y|\mu, \varphi)}{\partial \mu} = \frac{y - \mu}{\mu(1 + \mu \varphi)}
\]
\[
\frac{\partial^2 \ln p(y|\mu, \varphi)}{\partial \mu^2} = -\frac{y + \mu \varphi(3y - 2\mu)}{\mu^2(1 + \mu \varphi)^3},
\]
so the expected Hessian is
\[
E_{y|\mu, \varphi} \left[ \frac{\partial^2 \ln p(y|\mu, \varphi)}{\partial \mu^2} \right] = -\frac{1}{\mu(1 + \mu \varphi)^2}.
\]

Also,
\[
\frac{\partial \ln p(y|\mu, \varphi)}{\partial \varphi} = \frac{\mu^2 - \{1 + \mu [2 + \varphi(2 + \mu \varphi)]\} y + y^2}{(1 + \mu \varphi)^2(1 + \varphi y)}
\]
\[
\frac{\partial^2 \ln p(y|\mu, \varphi)}{\partial \varphi^2} = \frac{3\mu^2 y}{(1 + \mu \varphi)^2} - \frac{(y - 1)y^2}{(1 + \varphi y)^2} - \frac{2\mu(1 + \varphi y)}{(1 + \mu \varphi)^3},
\]
and
\[
E_{y|\mu, \varphi} \left[ \frac{\partial^2 \ln p(y|\mu, \varphi)}{\partial \varphi^2} \right] = \frac{\mu^3}{(1 + \mu \varphi)^2} - \frac{E_{y|\mu, \varphi} \left[ (y - 1)y^2 \right]}{(1 + \varphi y)^2},
\]
where the latter expectation can easily be obtained by simulation.

**Appendix D. Details for the beta regression model**

The density for a single observation \(y|\mu, \varphi \sim \text{Beta} [y|\varphi \mu, \varphi (1 - \mu)]\) is of the form
\[
p(y|\mu, \varphi) = \frac{\Gamma(\varphi)}{\Gamma(\varphi \mu)\Gamma(\varphi (1 - \mu))} y^{\varphi \mu - 1} (1 - y)^{\varphi (1 - \mu) - 1},
\]

\[
\frac{\partial \ln p(y|\mu, \varphi)}{\partial \mu} = \frac{y - \mu}{\mu(1 + \mu \varphi)}
\]
\[
\frac{\partial^2 \ln p(y|\mu, \varphi)}{\partial \mu^2} = -\frac{1}{\mu^2(1 + \mu \varphi)},
\]
and
\[
E_{y|\mu, \varphi} \left[ \frac{\partial^2 \ln p(y|\mu, \varphi)}{\partial \mu^2} \right] = \frac{1}{\mu^2(1 + \mu \varphi)^2}.
\]
with mean $\mu$ and variance $\mu(1 - \mu)/(1 + \varphi)$. We then have
\[
\frac{\partial \ln p(y|\mu, \varphi)}{\partial \mu} = \varphi \left\{ \kappa_0 [\varphi(1 - \mu)] - \kappa_0 (\varphi \mu) - \ln \frac{1 - y}{y} \right\}
\]
\[
\frac{\partial^2 \ln p(y|\mu, \varphi)}{\partial \mu^2} = -\varphi^2 (\kappa_1 [\varphi(1 - \mu)] + \kappa_1 (\varphi \mu)),
\]
which is also the expected Hessian. $\kappa_0(\cdot)$ and $\kappa_1(\cdot)$ are the digamma and trigamma functions.

Also,
\[
\frac{\partial \ln p(y|\mu, \varphi)}{\partial \varphi} = -\mu \ln \frac{1 - y}{y} + \ln(1 - y) + \kappa_0(\varphi) - \mu \kappa_0(\mu \varphi) + (\mu - 1) \kappa_0 [\varphi(1 - \mu)]
\]
\[
\frac{\partial^2 \ln p(y|\mu, \varphi)}{\partial \varphi^2} = \kappa_1 (\varphi) - \mu^2 \kappa_1 (\mu \varphi) - (\mu - 1)^2 \kappa_1 [\varphi(1 - \mu)],
\]
which is also the expected Hessian.